Mixture-of-Experts Policies with Imitation Guidance for Multitask Multi-Agent Reinforcement Learning

1st Yuan Wang
School of Artificial Intelligence, UCAS
Institute of Automation, CAS
Beijing, China
wangyuan2025@ia.ac.cn

2nd Zhiqiang Pu
School of Artificial Intelligence, UCAS
Institute of Automation, CAS
Beijing, China
zhiqiang.pu@ia.ac.cn

3rd Jinyuan Feng
School of Artificial Intelligence, UCAS
Institute of Automation, CAS
Beijing, China
fengjinyuan2022@ia.ac.cn

4th Xiaolin Ai
Institute of Automation, CAS
Beijing, China
xiaolin.ai@ia.ac.cn

5th Tenghai Qiu
Institute of Automation, CAS
Beijing, China
tenghai.qiu@ia.ac.cn

6th Xiwen Ma
Institute of Automation, CAS
Beijing, China
maxw2021@sjtu.edu.cn

***Abstract*—Multitask multi-agent reinforcement learning (MARL) aims to train general-purpose cooperative agents across diverse scenarios, but faces fundamental challenges stemming from the multi-modal nature of joint behavior distributions. Different tasks require distinct coordination patterns, posing dual challenges. First, the expressiveness challenge arises because a single shared policy lacks capacity to capture all modes, leading to negative transfer. Second, the mode coverage challenge emerges as standard mode-seeking optimization of reinforcement learning struggles to discover and maintain diverse coordination modes. Combining sparse Mixture-of-Experts (MoE) with generative adversarial imitation learning (GAIL), GEMS (GAIL-Enhanced MoE for multitask Specialization) is developed to address both challenges synergistically. MoE provides structured capacity by decomposing the policy into specialized experts, while GAIL delivers mode-covering guidance through expert demonstrations that embody diverse task-specific coordination patterns. MoE offers the expressiveness to represent multiple modes, while GAIL ensures effective mode coverage by accelerating expert differentiation and preventing mode collapse. Experiments on StarCraft Multi-Agent Challenge (SMAC) and Multi-Agent Particle Environment (MPE) demonstrate that GEMS significantly outperforms state-of-the-art methods in sample efficiency, final performance, and zero-shot generalization.**

***Index Terms*—Multitask reinforcement learning, multi-agent system, mixture of experts, imitative learning.**

I. Introduction

Multi-agent reinforcement learning (MARL) provides a principled framework for learning coordinated behaviors among agents under coupled dynamics and partial observability. Coordination is challenging due to the non-stationarity caused by simultaneously learning and evolving agents. Centralized Training with Decentralized Execution (CTDE) effectively alleviates this issue by leveraging global information during the training phase and maintaining policy distribution during the execution phase [1], [2]. Various MARL algorithms based on CTDE demonstrate strong performance in benchmarks such as StarCraft Multi-Agent Challenge (SMAC) [3], Multi-Agent Particle Environment (MPE) [4], and Multi-Agent MuJoCo [5], and are becoming increasingly significant in real-world applications, including game AI [6], football [7], and drone swarms [8].

Under the CTDE paradigm, a broad family of MARL algorithms has demonstrated remarkable performance in various complex multi-agent environments. These impressive performances are mostly achieved through single-task training mechanisms. HOA-Net [9] is designed to address observational heterogeneity in heterogeneous multi-agent systems by modeling observations as heterogeneous graphs and leveraging class-specific networks. A policy resonance approach [10] is proposed to solve the responsibility diffusion problem, refactoring joint agent policies while maintaining individual policy invariance. ACPPO [11] is an advantage-constrained policy gradient algorithm to integrate value-based and policy gradient methods, which can constrain joint and local advantages. When faced with multitask settings involving different scenarios, issues such as gradient interference and negative transfer often arise, which can reduce the training stability and data efficiency of general algorithms.

The task diversity in multitask MARL fundamentally induces a multi-modal joint behavior distribution [12], as different tasks require distinct coordination patterns. This multi-modality poses two complementary challenges that must be addressed simultaneously. The first challenge is the lack of expressiveness. A single shared policy has insufficient capacity to capture all coordination modes effectively, leading to negative transfer when task-specific patterns conflict. Improvements in certain tasks may come at the expense of others due to gradient interference across dissimilar task dynamics [13]. Hierarchical skill discovery approaches such as HMASD [14] alleviate this via task decomposition into modular skills, yet rely on policy-centric designs that fail to fully isolate diverse task dynamics. Sequence-modeling MARL methods such as MAT [15], which model multi-agent interactions as sequences, also struggle with this capacity limitation. The second challenge is the lack of mode coverage. Standard reinforcement learning

(RL) optimization is inherently mode-seeking, meaning policy gradients tend to reinforce already-discovered coordination patterns while struggling to maintain or explore alternative modes [16]. In multitask settings, this causes the policy to collapse to a dominant mode, exhibit catastrophic forgetting across tasks [17], or fail to discover certain coordination patterns entirely during exploration. These dual challenges require not only sufficient model capacity, but also effective guidance to ensure all coordination modes are discovered and retained.

The Mixture of Experts (MoE) [18], [19] architecture addresses the expressiveness challenge by decomposing the policy into specialized experts with a routing function. This design has proven effective in large language models, where different inputs benefit from different sub-models [20]. In multitask MARL, experts can specialize in different coordination patterns while the routing function selects appropriate experts based on task context, thereby reducing gradient interference. Early MoE methods like CARE [21] introduced context-aware representations to differentiate tasks in shared policies, while PaCo [22] combined shared and task-specific parameters to boost cross-task transfer. M3W [23] applied MoE to world models, enhancing generalization by leveraging task dynamics similarity. However, MoE alone cannot fully address the mode coverage challenge. Without explicit guidance, expert differentiation remains difficult. The mode-seeking nature of RL may cause multiple experts to converge toward the same dominant coordination pattern, underutilizing the model capacity and failing to cover the full spectrum of task-specific modes.

To address above challenges, a novel framework named GEMS (**G**AIL-**E**nhanced **M**oE for multitask **S**pecialization) is developed. GEMS combines the structured capacity of MoE with the mode-covering guidance of generative adversarial imitation learning (GAIL) [24]. The actor is parameterized as a sparseMoE to provide expressiveness for multi-modal coordination patterns. To ensure effective mode coverage, expert demonstrations collected from strong policies across different tasks are incorporated via GAIL. These demonstrations naturally embody diverse task-specific coordination modes [12], providing a mode-covering signal that accelerates expert specialization and prevents mode collapse. MoE provides the capacity to express multiple modes through specialized experts, while GAIL ensures all modes are discovered, learned, and maintained throughout training by exposing experts to diverse behavioral exemplars. This combination enables robust multitask MARL policies that generalize across diverse coordination scenarios. The main contributions of this work are summarized as follows:

- A sparseMoE actor is introduced for multitask MARL, enabling structured policy decomposition that addresses the expressiveness challenge by allowing experts to specialize in distinct coordination patterns while mitigating negative transfer across tasks.
- GAIL-based imitation guidance is incorporated into the multitask MARL training pipeline to address the mode coverage challenge, accelerating expert differentiation by leveraging diverse expert demonstrations as behavioral exemplars for different tasks.
- Extensive experiments conducted on SMAC and MPE demonstrate that GEMS outperforms strong baseline models in sample efficiency, final performance, and zero-shot generalization capability, with mechanistic analysis confirming effective task-dependent expert specialization.

II. Preliminaries

A. Problem Formulation

Cooperative MARL problems can be modeled as a decentralized partially observable Markov decision process (Dec-POMDP) [25], which is defined as the tuple $\mathcal{M} = \langle \mathcal{I}, \mathcal{S}, \{\mathcal{O}^i\}_{i\in\mathcal{I}}, \{\mathcal{A}^i\}_{i\in\mathcal{I}}, P, O, r, \gamma \rangle$, where $\mathcal{I} = \{1, \ldots, N\}$ is the set of agents, $\mathcal{S}$ is the global state space, $\mathcal{A}^i$ and $\mathcal{O}^i$ denote the action and observation spaces of agent i, $P(s' \mid s, \mathbf{a})$ is the state transition probability, $O(o^i \mid s, i)$ is the observation function, r is the shared reward, and $\gamma \in [0, 1)$ is the discount factor. In multitask MARL, the objective is to train a model that performs well across a finite set of tasks $\mathcal{T} = \{\mathcal{M}_k\}$, where each task $\mathcal{M}_k$ is a Dec-POMDP. During training, a task τ is sampled from the task distribution $p(\tau)$ over $\mathcal{T}$, and the corresponding environment is denoted as $\mathcal{M}_\tau$.

B. MoE Mechanism

MoE consists of K expert networks and a gating function that routes each input to a subset of experts. For the input representation x, the gate outputs a weight vector $g(x) \in \mathbb{R}^K$. The MoE output is computed as the following weighted mixture of expert outputs:

$$y(x) = \sum_{k=1}^{K} g_k(x)\, f_k(x) \tag{1}$$

where $f_k(\cdot)$ is the k-th expert and $g_k(x)$ is corresponding weight. MoE models are commonly categorized into SoftMoE [26] and SparseMoE [18]. In SoftMoE, all experts are mixed for every input, which is straightforward but the computational burden increases sharply as the number of experts increases. Inversely, SparseMoE activates only a subset of experts for each input, achieving greater model capacity with lower computational cost. Beyond efficiency, sparse routing naturally reduces gradient interference by directing different coordination modes to different experts, and encourages expert specialization since each expert receives focused training signals from a subset of inputs. Given these advantages for multi-modal multitask learning, sparseMoE is adopted in this work.

C. Imitation Learning

Imitation learning aims to learn a policy from expert demonstrations, providing a practical strategy when designing reward functions is difficult or when exploration is inefficient. Beyond directly imitating expert actions, inverse RL seeks to infer a reward function that explains expert behavior and then optimizes a policy under the inferred reward. Adversarial imitation learning leverages demonstrations by training a

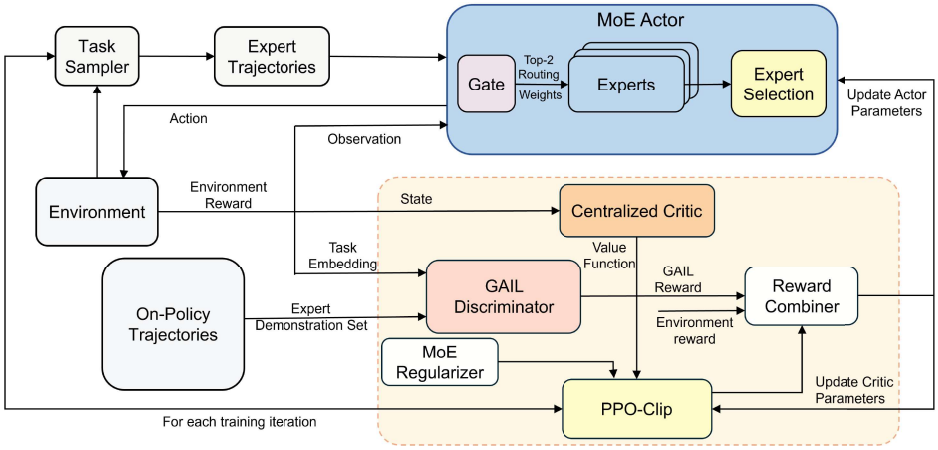

Fig. 1. Overall framework of GEMS. Left: a task τ is sampled and agents interact with the environment $\mathcal{M}_\tau$ using task embedding e_τ. Top-right (blue): decentralized execution via shared MoE actor, where a gate g_ω routes inputs to top-2 experts and outputs a mixed action distribution. Bottom-right (orange): centralized training components including the critic V_ψ, team-level GAIL discriminator D_ϕ, reward combiner, and MAPPO update with load-balancing regularization.

discriminator to distinguish expert trajectories from policy-generated trajectories, while updating the policy to make its behavior indistinguishable from the expert. The common formulation is the following min-max objective:

$$\min_{\pi} \max_{D} \mathbb{E}_{\tau_{\text{traj}}\sim\mathcal{D}_E}\big[\log D(\tau_{\text{traj}})\big] + \mathbb{E}_{\tau_{\text{traj}}\sim\pi}\big[\log\big(1-D(\tau_{\text{traj}})\big)\big] \tag{2}$$

where $\mathcal{D}_E$ denotes the expert demonstration set, τ_{traj} denotes the trajectory, and $D(\cdot)$ is a discriminator that provides a learning signal for policy optimization.

III. Framework Overview

Multitask cooperative MARL is investigated under the CTDE paradigm. During training, tasks are repeatedly sampled from $p(\tau)$ and on-policy trajectories are collected by interacting with $\mathcal{M}_\tau$. All agents share a learnable task embedding e_τ to facilitate task-aware decision-making. The framework overview is shown in Fig. 1.

GEMS is built upon Multi-Agent Proximal Policy Optimization (MAPPO) [27] with two key extensions addressing complementary challenges. First, to address the expressiveness challenge, the actor is parameterized as a sparseMoE, decomposing a single shared policy into multiple expert sub-policies. A task-conditioned gate routes each input to a subset of experts, and the final action distribution is computed as a weighted mixture of the selected experts. This structured decomposition provides the capacity to represent multi-modal coordination patterns while allowing experts to specialize in distinct modes, thereby mitigating gradient interference across tasks.

Second, to address the mode coverage challenge, GAIL-based imitation guidance is introduced. A team-level discriminator distinguishes expert demonstrations from policy-generated samples, providing a mode-covering signal that ensures diverse coordination patterns are discovered and maintained. The discriminator output is transformed into a shaping reward combined with environment rewards via a time-dependent weight $\lambda(t)$ that is annealed to emphasize mode-covering guidance early while allowing task-specific environment rewards to dominate later.

Training alternates between collecting on-policy trajectories, updating the discriminator, computing shaped rewards, and updating the MoE actors and centralized critic. The following section details each component.

IV. Method

This section presents the technical details of GEMS. First, the MAPPO algorithm with a centralized critic is introduced as the base framework for multitask MARL. Then, the sparse-MoE actor is described, which addresses the expressiveness challenge by decomposing the policy into specialized experts. Finally, the GAIL-based imitation shaping mechanism is presented, which provides mode-covering guidance to ensure diverse coordination patterns are discovered and maintained throughout training.

A. MAPPO with Centralized Critic for Multitask MARL

MAPPO is adopted as the base algorithm under the CTDE paradigm for multitask training over a task distribution $\tau \sim p(\tau)$. Each agent executes a decentralized policy $a_t^i \sim \pi_\theta(\cdot \mid o_t^i, e_\tau)$ using its local observation and task embedding e_τ, where θ denotes shared actor parameters.

A centralized state-value critic $V_\psi(c_t, e_\tau)$ is employed during training, where c_t is available only at training time. Advantages are computed via generalized advantage estimator as follows:

$$\begin{aligned}\hat{A}_t &= \sum_{l=0}^{T-t-1} (\gamma\lambda_{\text{GAE}})^l \delta_{t+l} \\ \delta_t &= r_t + \gamma V_\psi(c_{t+1}, e_\tau) - V_\psi(c_t, e_\tau)\end{aligned} \tag{3}$$

where γ is the discount factor and λ_{GAE} controls the bias–variance trade-off. The same advantage estimate is used for all agents due to the shared reward structure.

The PPO clipped objective for agent i is

$$L^i_{\text{clip}}(\theta) = \mathbb{E}_t\Big[\min\Big(\rho^i_t\,\hat{A}_t,\ \text{clip}(\rho^i_t,\,1-\epsilon,\,1+\epsilon)\,\hat{A}_t\Big)\Big] \quad (4)$$

where $\rho^i_t = \pi_\theta(a^i_t \mid o^i_t, e_\tau)/\pi_{\theta_{\text{old}}}(a^i_t \mid o^i_t, e_\tau)$. The averaged objective across agents $L_{\text{clip}}(\theta) = \frac{1}{N}\sum_{i=1}^{N} L^i_{\text{clip}}(\theta)$ is optimized jointly with the critic loss via following MSE regression and entropy bonus:

$$L_{\text{MAPPO}}(\theta,\psi) = -L_{\text{clip}}(\theta) + c_V\,L_V(\psi) - \beta_{\text{ent}}\,L_{\text{ent}}(\theta). \quad (5)$$

In the multitask setting, trajectories from sampled tasks are aggregated and optimized over mixed batches to maximize $\mathbb{E}_{\tau\sim p(\tau)}[J_\tau(\pi_\theta)]$.

B. MoE Actor for Multitask MARL

Different tasks induce distinct coordination patterns, resulting in multi-modal joint behavior distributions. This poses the expressiveness challenge. A single shared policy lacks sufficient capacity to capture all coordination modes effectively. An MoE architecture addresses this by decomposing the policy into K specialized experts with a gating network, providing the representational capacity for multi-modal behaviors.

For agent i, given observation o^i_t and task embedding e_τ, the gating network $g_\omega(\cdot)$ outputs routing scores

$$\mathbf{s}^i_t = g_\omega(o^i_t, e_\tau) \in \mathbb{R}^K \quad (6)$$

which can convert to probabilities via softmax $\mathbf{p}^i_t = \text{softmax}(\mathbf{s}^i_t)$. To maintain computational efficiency, only the top-2 experts are activated as follows:

$$\mathcal{K}^i_t = \text{Top2}(\mathbf{p}^i_t), \quad w^i_{t,k} = \frac{p^i_{t,k}}{\sum_{j\in\mathcal{K}^i_t} p^i_{t,j}} \quad (k \in \mathcal{K}^i_t). \quad (7)$$

The final policy is a mixture of selected expert action distributions, represented by

$$\pi_\theta(a^i_t \mid o^i_t, e_\tau) = \sum_{k\in\mathcal{K}^i_t} w^i_{t,k}\,\pi_{\theta_k}(a^i_t \mid o^i_t, e_\tau) \quad (8)$$

where $\theta = \{\omega, \theta_1, \ldots, \theta_K\}$. This distribution-level mixture provides the capacity to express diverse coordination modes while allowing experts to specialize in distinct task-specific patterns.

All agents share the same gate and experts but condition on local observations, enabling diverse per-agent behaviors. To encourage balanced expert utilization complementing the mode-covering guidance from GAIL, a load-balancing regularizer is applied

$$L_{\text{bal}}(\omega) = \sum_{k=1}^{K} \bar{p}_k \log(\bar{p}_k K) \quad (9)$$

where $\bar{p}_k = \mathbb{E}_{t,i}[p^i_{t,k}]$ denotes average routing probability. This KL divergence term encourages diverse expert utilization.

The MoE actor is optimized via the PPO objective (4) with action probability computed by the equation (8). The importance sampling ratio becomes

$$\rho^i_t(\theta) = \frac{\sum_{k\in\mathcal{K}^i_t} w^i_{t,k}\,\pi_{\theta_k}(a^i_t \mid o^i_t, e_\tau)}{\sum_{k\in\mathcal{K}^i_t} w^{i,\text{old}}_{t,k}\,\pi_{\theta^{\text{old}}_k}(a^i_t \mid o^i_t, e_\tau)} \quad (10)$$

where $\mathcal{K}^i_t$ is determined by the old gate $g_{\omega_{\text{old}}}$ during rollout collection and remains fixed during policy updates. Gradients jointly update the gate and selected experts, learning both routing decisions and expert specializations. By allocating different coordination modes to different experts, the MoE structure addresses the expressiveness challenge and mitigates gradient interference across tasks.

C. GAIL-based Imitation Shaping

While MoE addresses the expressiveness challenge by providing capacity to represent multiple coordination modes, the mode coverage challenge remains. Standard RL optimization is mode-seeking, tending to reinforce already-discovered patterns while struggling to maintain or explore alternative modes. In multitask MARL, this causes experts to converge toward similar dominant patterns despite having capacity for diverse specialization. To address this, GAIL is incorporated to provide mode-covering guidance through expert demonstrations collected from different tasks, which naturally embody diverse coordination modes. Notably, the GAIL signal is used only during training and does not affect decentralized execution.

Expert demonstrations $\mathcal{D}_E$ are collected from strong policies across the training task set, providing diverse behavioral exemplars that embody task-specific coordination modes. At each time step t, a demonstration sample comprises team-level information, represented by

$$x^E_t = (c^E_t, a^E_t), \quad a^E_t = (a^{1,E}_t, \ldots, a^{N,E}_t). \quad (11)$$

Similarly, policy rollouts are sampled as

$$x^\pi_t = (c_t, a_t), \quad a_t = (a^1_t, \ldots, a^N_t). \quad (12)$$

Using centralized observations and joint actions enables the discriminator to capture multi-agent coordination patterns, which are the fundamental building blocks of distinct behavioral modes in cooperative multitask MARL.

A discriminator $D_\phi : \mathcal{X} \to [0,1]$ is trained to distinguish expert and policy samples by minimizing

$$L_D(\phi) = -\mathbb{E}_{x\sim\mathcal{D}_E}\big[\log D_\phi(x)\big] - \mathbb{E}_{x\sim\mathcal{D}_\pi}\big[\log(1-D_\phi(x))\big]. \quad (13)$$

The discriminator output is transformed into an imitation reward

$$r^{\text{gail}}_t = \log D_\phi(x^\pi_t) \quad (14)$$

which provides a mode-covering signal by rewarding behaviors that match the diverse demonstrations. This is combined with the environment reward

$$r_t = r^{\text{env}}_t + \lambda(t)\,r^{\text{gail}}_t \quad (15)$$

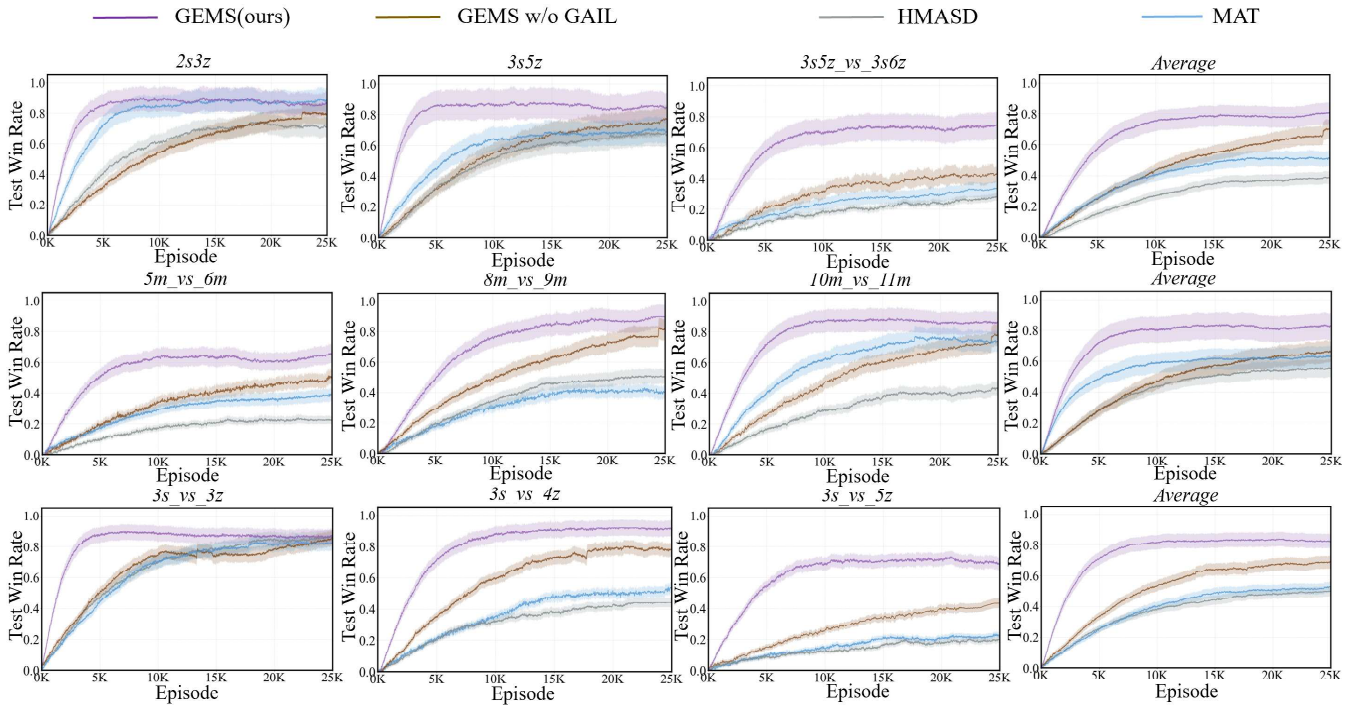

Fig. 2. Multitask performance on SMAC. Each row corresponds to one task set. The first three columns show per-task win rates and the fourth column shows the mean win rate across tasks.

where $\lambda(t)$ is a time-dependent weight annealed over training to emphasize mode-covering guidance early while allowing task-specific environment rewards to dominate later, balancing exploration of diverse modes with task optimization.

Since tasks are fully cooperative with shared rewards, the team-level GAIL signal applies uniformly to all agents at time step t, avoiding per-agent inconsistency and aligning with centralized value estimation. During decentralized execution, the GAIL signal is absent, preserving decentralized decision-making based solely on (o_t^i, e_τ).

Training alternates between the following steps: (i) sampling tasks from $p(\tau)$ and collecting on-policy trajectories via decentralized execution; (ii) updating the discriminator to distinguish expert from policy samples constructed from team-level pairs (c_t, a_t); (iii) computing shaped rewards via the equation (15); (iv) updating MoE actors and centralized critic via the objective (5) together with the load-balancing regularizer (9). Multiple optimization epochs per rollout follow standard PPO practice.

V. Experiment

In this section, comprehensive experiments are presented to evaluate GEMS. The experimental setup and baselines are first introduced. Then, multitask performance and sample efficiency are compared against state-of-the-art methods. Finally, mechanistic analyses of mode coverage and expert specialization are provided, followed by zero-shot generalization evaluation.

A. Experimental Setup

GEMS is evaluated on two standard multi-agent benchmarks, SMAC and MPE, under multitask training settings.

For SMAC, three multitask sets are constructed as follows:

- **Set1:** {5m_vs_6m, 8m_vs_9m, 10m_vs_11m};
- **Set2:** {2s3z, 3s5z, 3s5z_vs_3s6z};
- **Set3:** {3s_vs_3z, 3s_vs_4z, 3s_vs_5z}.

Tasks within each set share similar unit types and dynamics while differing in opponent strength.

For MPE, two multitask sets are constructed as follows:

- **Spread:** {spread_4N, spread_6N, spread_8N}, where the number of agents equals the number of landmarks;
- **Tag:** {tag_3v1, tag_5v2, tag_7v3}, with 8 obstacles and prey speed and acceleration set to twice that of the controlled agents.

These tasks induce distinct coordination modes due to scale variations.

At each episode, a task τ is sampled uniformly from the task set, and all agents share a learnable task embedding e_τ. Trajectories collected from different tasks are aggregated for the algorithm updates. Tasks with varying agent numbers are handled by padding observations and masking invalid agents or actions.

Imitation guidance is provided by a GAIL-based discriminator trained on enough expert episodes collected from a strong MAPPO policy checkpoint. Expert trajectories are generated deterministically on GPU with a fixed random seed and are used only during training.

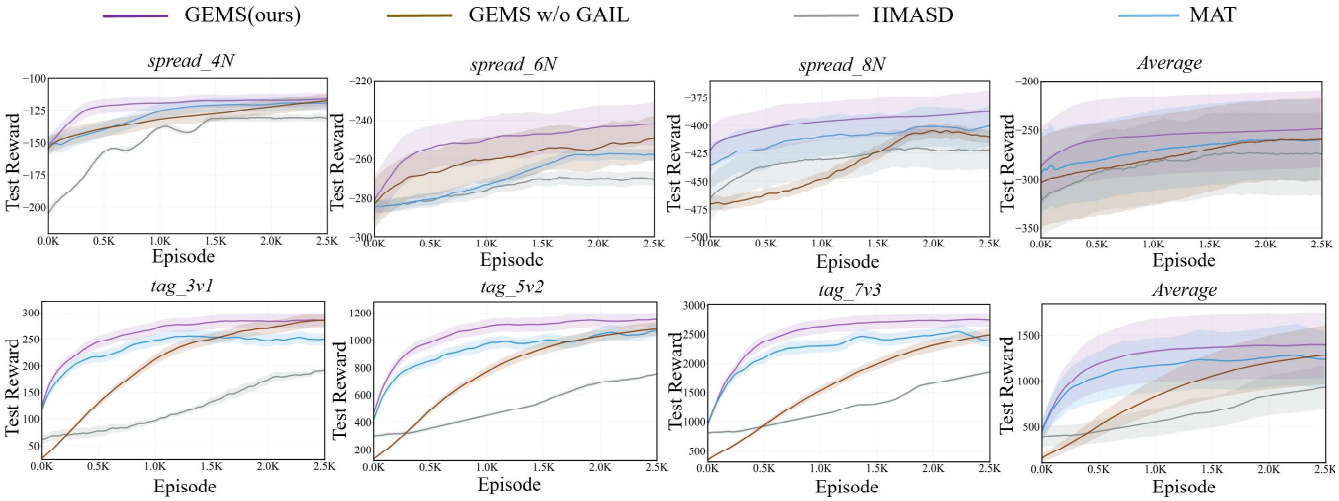

Fig. 3. Multitask performance on MPE. Each row corresponds to one task set. The first three columns show per-task rewards and the fourth column shows the mean reward across tasks.

GEMS is compared with strong multi-agent baselines, including MAT [15] and HMASD [14]. In addition, an ablation variant GEMS w/o GAIL is included to isolate the effect of imitation guidance within the GEMS framework.

B. Multitask Performance and Sample Efficiency

GEMS is evaluated on multitask settings in SMAC and MPE, focusing on overall performance, sample efficiency, and training stability. In SMAC, win rates are reported. In MPE, episode return rewards are reported when applicable. Results are summarized in Fig. 2 and Fig. 3. All curves are averaged through 5 random seeds. Solid lines denote the mean performance and shaded areas indicate one standard deviation. For visualization, a sliding-window moving average with a window size of 20 episodes is applied.

Fig. 2 reports the win rates on three SMAC multitask sets. Each row corresponds to one task set, where the first three columns show per-task performance and the fourth column shows the mean win rate across tasks within the same set. The mean win rate is computed by averaging the win rates of all tasks in the set at each episode, providing a holistic measure of multitask performance.

Across all three task sets, GEMS consistently achieves higher average win rates and faster convergence compared to MAT and HMASD. Notably, the performance gap is more pronounced on harder tasks, such as 10m_vs_11m and 3s5z_vs_3s6z, indicating improved robustness under increasing task difficulty. Compared to the GEMS w/o GAIL variant, the full GEMS method exhibits more stable early learning and higher final performance. This performance advantage is attributed to effective mode coverage. GAIL-based guidance ensures that diverse task-specific coordination patterns are discovered and maintained rather than collapsing to a single dominant mode, enabling the MoE policy to leverage its full capacity for expressing multi-modal behaviors.

Fig. 3 presents the results on two MPE multitask sets. Each row corresponds to one task set, and each column follows the same layout as in SMAC. Due to differences in task objectives and scale, the reward ranges vary across task sets. Therefore, each row uses an independent y-axis scale. The fourth column reports the mean reward over tasks within the same set.

GEMS outperforms the baselines in both the spread and tag task sets. In particular, as the number of agents increases, coordination patterns become more complex and multi-modal, posing greater challenges for mode coverage. GEMS maintains stable learning and achieves higher average rewards, while MAT and HMASD exhibit slower convergence or degraded performance on larger-scale tasks. Similar to SMAC, removing GAIL-based mode-covering guidance from the MoE actor leads to significantly degraded performance, especially in early training stages where RL mode-seeking nature would otherwise cause the policy to prematurely converge to suboptimal coordination modes without exploring the full spectrum of effective patterns.

A GEMS w/o MoE baseline is intentionally not included, as the two components address complementary challenges. Firstly, MoE provides the expressiveness to represent multiple coordination modes, while GAIL provides the mode-covering guidance to discover and maintain them. Secondly, without MoE structured capacity, a single shared policy remains fundamentally limited in expressing the multi-modal coordination patterns induced by multiple tasks, regardless of the quality of imitation guidance. Hence, the ablation focuses on isolating the effect of mode-covering guidance within the MoE framework, demonstrating that both capacity and coverage are necessary for effective multitask MARL.

Overall, these results demonstrate that GEMS achieves multitask performance and sample efficiency across both SMAC and MPE. The performance gains validate the synergistic combination of MoE expressiveness and GAIL mode-covering

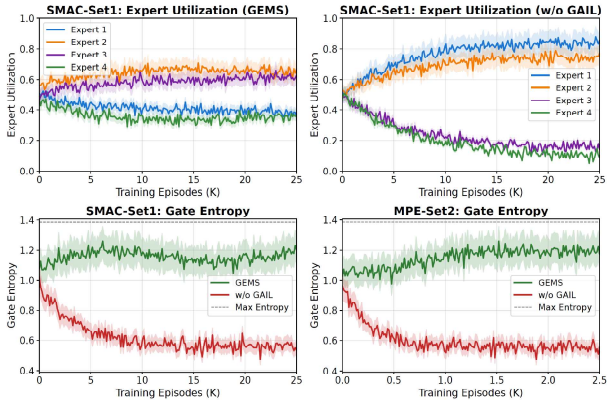

Fig. 4. Mode coverage analysis on SMAC-Set1 and MPE-Set2. Expert utilization and gate entropy during training are reported for GEMS and GEMS w/o GAIL. Higher entropy and balanced utilization indicate effective mode coverage.

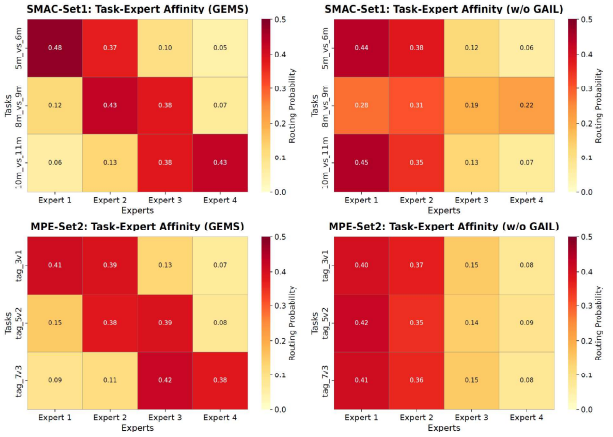

Fig. 5. Task–expert affinity heatmaps on SMAC-Set1 and MPE-Set2. Each entry shows $\bar{p}_{\tau,k} = \mathbb{E}_{(t,i)\sim\tau}[p^i_{t,k}]$, the average gate probability assigned to expert k on task τ.

guidance. By ensuring that diverse coordination patterns are discovered, learned, and maintained throughout training, GEMS can addresses both the capacity and mode coverage challenges inherent in multitask MARL.

C. Mode Coverage and Expert Specialization Analysis

To validate that GAIL provides mode-covering guidance enabling expert specialization, this subsection analyzes whether GAIL maintains diverse expert utilization throughout training and whether experts develop structured task-dependent specialization patterns.

Expert utilization $u_k = \mathbb{E}_{t,i}[\mathbb{I}(k \in \text{Top2}(p^i_t))]$ is used to measure whether all experts remain active. Gate entropy $H(p) = -\mathbb{E}_{t,i} \sum_{k=1}^{K} p^i_{t,k} \log p^i_{t,k}$ is employed to quantify routing diversity, where high entropy indicates mode coverage and low entropy suggests mode collapse. Fig. 4 compares GEMS and GEMS w/o GAIL on SMAC-Set1 and MPE-Set2 with 4 experts. Without GAIL, gate entropy drops sharply and expert utilization becomes highly imbalanced, indicating the mode-seeking nature of RL causes convergence to a single dominant pattern. GEMS maintains consistently higher entropy and balanced utilization, demonstrating effective mode coverage.

To examine task-dependent specialization, task–expert affinity $\bar{p}_{\tau,k} = \mathbb{E}_{(t,i)\sim\tau}[p^i_{t,k}]$ is computed. Fig. 5 shows clear task-dependent patterns under GEMS with 4 experts. Different tasks favor distinct expert subsets, indicating functionally differentiated coordination modes. Compared to GEMS w/o GAIL, GEMS produces more structured routing preferences, confirming that expert demonstrations provide effective mode-covering signals for meaningful specialization.

Therefore, without GAIL, MoE policies exhibit mode collapse due to the mode-seeking nature. Furthermore, GAIL effectively addresses mode coverage by maintaining diverse utilization and structured specialization. This mechanistc understanding explains the performance gains in Section V-B.

TABLE I
ZERO-SHOT GENERALIZATION ON HELD-OUT TASKS (NO FINE-TUNING). TRAIN SETS: S1: {5M_VS_6M, 8M_VS_9M}; S2: {3S_VS_3Z, 3S_VS_4Z}; M1: {SPREAD_4N, SPREAD_6N}; M2: {TAG_3V1, TAG_5V2}. MEAN ± STD OVER 5 SEEDS ARE REPORTED.

ID	Test	GEMS	w/o GAIL	MAT	HMASD
S1	10m_vs_11m (Win)	0.32±0.07	0.26±0.11	0.27±0.08	0.21±0.15
S2	3s_vs_5z (Win)	0.23±0.05	0.17±0.06	0.21±0.11	0.19±0.05
M1	spread_8N (Ret.)	-410±14	-430±21	-415±12	-412±17
M2	tag_7v3 (Ret.)	2056±122	1981±145	2003±101	2010±134

D. Zero-Shot Generalization

To evaluate generalization to unseen tasks, models are trained on task subsets and evaluated on held-out tasks without fine-tuning. In SMAC, training uses {5m_vs_6m, 8m_vs_9m} with evaluation on 10m_vs_11m, and {3s_vs_3z, 3s_vs_4z} with evaluation on 3s_vs_5z. In MPE, training uses {tag_3v1, tag_5v2} with evaluation on tag_7v3. Table I shows GEMS achieves the zero-shot performance across all held-out tasks with lower variance, indicating improved robustness to task shifts.

The improved zero-shot capability stems from effective mode coverage. MoE provides capacity to learn diverse coordination modes, while GAIL ensures these modes are discovered and retained. By learning diverse yet structured patterns across training tasks, experts can flexibly adapt to related unseen tasks. This is consistent with Section V-C, where GEMS preserves multiple coordination modes through balanced utilization and structured routing. Methods lacking mode-covering guidance or capacity struggle to generalize.

VI. CONCLUSION

This paper proposes GEMS, a framework addressing the dual challenges of expressiveness and mode coverage in multitask MARL through the synergistic combination of sparse-MoE and GAIL-based imitation guidance. MoE provides the capacity to represent multi-modal coordination patterns via

specialized experts, while GAIL ensures diverse modes are discovered and maintained through mode-covering guidance from expert demonstrations. Experiments on SMAC and MPE demonstrate improvements in sample efficiency, final performance, and zero-shot generalization over baselines. Future work can explore Sim-to-Real transfer of the learned multitask policies, leveraging GEMS robust expert decomposition to bridge the reality gap for real-world multi-agent systems.

ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China under Grant No. 62503472 and Grant 62322316, the Young Scientists Foundation of CSAA (Guidance Navigation and Control, GNC) under Grant CSAA-YSF2025-GNC-08, and the CAS Project for Young Scientists in Basic Research (No. YSBR-123).

REFERENCES

[1] J. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson, "Counterfactual multi-agent policy gradients," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[2] T. Rashid, M. Samvelyan, C. S. De Witt, G. Farquhar, J. Foerster, and S. Whiteson, "Monotonic value function factorisation for deep multi-agent reinforcement learning," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 7234–7284, 2020.

[3] O. Vinyals, I. Babuschkin, W. M. Czarnecki, et al., "Grandmaster level in StarCraft II using multi-agent reinforcement learning," *Nature*, vol. 575, pp. 350–354, 2019.

[4] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," *Advances in Neural Information Processing Systems*, vol. 30, pp. 6382–6393, 2017.

[5] B. Peng, T. Rashid, C. A. S. de Witt, P.-A. Kamienny, P. H. S. Torr, W. Bohmer, and S. Whiteson, "FACMAC: Factored multi-agent centralised policy gradients," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12208–12221, 2021.

[6] B. Ellis, J. Cook, S. Moalla, M. Samvelyan, M. Sun, A. Mahajan, J. N. Foerster, and S. Whiteson, "SMACv2: An improved benchmark for cooperative multi-agent reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[7] K. Kurach, A. Raichuk, P. Stanczyk, M. Zajac, O. Bachem, L. Espeholt, C. Riquelme, D. Vincent, M. Michalski, O. Bousquet, and S. Gelly, "Google research football: A novel reinforcement learning environment," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 4, pp. 4501–4510, 2020.

[8] Y. -J. Chen, D. -K. Chang and C. Zhang, "Autonomous tracking using a swarm of UAVs: A constrained multi-agent reinforcement learning approach," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 11, pp. 13702–13717, 2020.

[9] T. Hu, X. Ai, Z. Pu, T. Qiu, and J. Yi, "Heterogeneous observation aggregation network for multi-agent reinforcement learning," in *2024 International Joint Conference on Neural Networks*, pp. 1–9, 2024.

[10] Q. Fu, T. Qiu, J. Yi, Z. Pu, X. Ai, and W. Yuan, "A policy resonance approach to solve the problem of responsibility diffusion in multiagent reinforcement learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 5, pp. 9621–9635, 2025.

[11] W. Li, Y. Zhu, and D. Zhao, "Advantage constrained proximal policy optimization in multi-agent reinforcement learning," in *2023 International Joint Conference on Neural Networks*, pp. 1–8, 2023.

[12] H. Mandlekar, D. Xu, J. Wong, S. Nasiriany, C. Wang, R. Kulkarni, F.-F. Li, S. Savarese, Y. Zhu, and R. Martín-Martín, "What matters in learning from offline human demonstrations for robot manipulation," in *Conference on Robot Learning*, pp. 1678–1690, 2022.

[13] J. Feng, M. Chen, Z. Pu, T. Qiu, J. Yi, and J. Zhang, "Efficient multitask reinforcement learning via task-specific action correction," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 17, no. 5, pp. 1110–1124, 2025.

[14] M. Yang, Y. Yang, Z. Lu, W. Zhou, and H. Li, "Hierarchical multi-agent skill discovery," *Advances in Neural Information Processing Systems*, 2023.

[15] M. Wen, J. G. Kuba, R. Lin, W. Zhang, Y. Wen, J. Wang, and Y. Yang, "Multi-agent reinforcement learning is a sequence modeling problem," *Advances in Neural Information Processing Systems*, 2022.

[16] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *International Conference on Machine Learning*, pp. 1861–1870, 2018.

[17] K. Khetarpal, M. Riemer, I. Rish, and D. Precup, "Towards continual reinforcement learning: A review and perspectives," *Journal of Artificial Intelligence Research*, vol. 75, pp. 1401–1476, 2022.

[18] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," in *International Conference on Learning Representations*, 2017.

[19] X. Chen, M. Ha, Z. Lan, J. Zhang, and J. Li, "MoBE: Mixture-of-basis-experts for compressing MoE-based LLMs," in *International Conference on Learning Representations*, 2026.

[20] N. Du, Y. Huang, A. M. Dai, et al., "GLaM: Efficient scaling of language models with mixture-of-experts," in *International Conference on Machine Learning*, 2021.

[21] A. R. Featherstone, A. Srinivas, M. Namkoong, and S. Levine, "Context-aware representation learning for multi-task reinforcement learning," in *International Conference on Machine Learning*, 2020.

[22] Y. Liu, A. Rajeswaran, S. Goyal, M. Bornstein, J. Schneider, and P. Abbeel, "Parameter composition for multi-task reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[23] Z. Zhao, Z. Zhao, K. Xu, Y. Fu, J. Chai, Y. Zhu, and D. Zhao, "Learning and planning multi-agent tasks via an MoE-based world model," in *Thirty-ninth Conference on Neural Information Processing Systems*, 2025.

[24] J. Ho and S. Ermon, "Generative adversarial imitation learning," *Advances in Neural Information Processing Systems*, pp. 4565–4573, 2016.

[25] M. L. Puterman, "Markov decision processes," *Handbooks in Operations Research and Management Science*, vol. 2, pp. 331–434, 1990.

[26] J. Puigcerver, C. Riquelme, B. Mustafa, and N. Houlsby, "From sparse to soft mixtures of experts," in *International Conference on Learning Representations*, 2024.

[27] C. Yu, A. Velu, E. Vinitsky, J. Gao, Y. Wang, A. Bayen, and Y. Wu, "The surprising effectiveness of PPO in cooperative multi-agent games," *Advances in Neural Information Processing Systems*, vol. 14, 2022.

[28] J. Schulman, P. Moritz, S. Levine, M. I. Jordan, and P. Abbeel, "High-dimensional continuous control using generalized advantage estimation," in *International Conference on Learning Representations*, 2016.